

SEQUENCE [L1] LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT:

- (A) NAME: Institut National de la Santé et de la
Recherche Médicale
(B) STREET: 101 rue de Tolbiac
(C) CITY: Paris
(E) COUNTRY: FRANCE
(F) POSTAL CODE (ZIP): 75013

(ii) TITLE OF INVENTION: oligophrenin 1 gene and protein

(iii) NUMBER OF SEQUENCES: 27

(iv) COMPUTER READABLE FORM:

- (A) MEDIUM TYPE: Floppy disk
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: PC-DOS/MS-DOS
(D) SOFTWARE: PatentIn Release #1.0, Version #1.25 (EPO)

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1650 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Homo sapiens

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

TTTTTCAACC CATGGTGCAT TTAAAGCTCA TTTTGGACAT TTTCCCATAA ATTAACCTTA
60

AAGATAAAAA GAGTAAGAAA CAAACTTTC CCCTGAGATG TGGCTACTTA TTTATTTTCA
120

GAGGGCGTTT TCTCATGCGC ATTCTGCTAT ATAGTGTAGT GGTCAAGAGC ACTACCTCTA
180

GAGCCAGCCA GGCTGGGCTC AAGTTCAAGT GCTGCCATTT AACTAGCTGT TTGTCCTTCG
240

GCAAGTCACT TAAACTCTCT TTGACCCAGC TTCTCCATCT TTAAATGGGT ATAATAATAA
300

AACCATCCTC ATAGGGTTGT TTTGAAGATT AGTGAGATGG GCGATAGGTT GTGTGGTGGG
360

TAGAATAATG TTCCTCCTT CACAGATGTC CATGTCCTGT CCTGAAACCT GTGGCTACGT
420

09504422 112100

TATCTTATGT GGCAAAATGA AATTTGCAGA TGTGATTAAG GATGATGTGA TGGGGGAGAT
480

TATCCTGTGG ACCCAGTGTA ATCATAAGGG TCCTTAAGGG GAGGCATGAG AATCACAGTG
540

ATGTAGCATG AGAAAGACTT GACTGGCCAG TGCTAGCTTT AAAGACGGAG GAAGGGAGCT
600

ATGAGCCAAA GGATGTGAGC AGCTTGAGCC GGAAAAGGTA AGGAACCAGA TTCTACCCCT
660

AGACTATCTG AAAATGAACA CAACCCTTCT AACATCTTGA TTTGATCCCA GTGAGGCCCA
720

TTTCTGATTT CTGCCTTCTG GAACTGTAGA TTATACTTT ATGCTGATTT AGGACATCAA
780

GTTTGTGGTA ATTTGTCACA GCAGCAGCAA CAGGACCTAA TTCAGGTGGG TCAGCCACTA
840

TTATTCTATT TACAACCCCC TTCCGTTCTG GCTTTCTTCA CAGGGCAATC TAATGTAAAC
900

ACTTCTGCAG AATTAAAGGG ATCTCTTTGC CCACGCACAA ACTATTTTTT AGTTTTTCGA
960

GCCTCTACCT ACCCTTAGTC TCAAGACAGC TGCCTTTAAG GTAAGCTGCA AGACAATTTA
1020

TAACATCCTA TTCTTAGACT TTCCACTTAC ACATGGAAAA TCATATGCAT TATCGGCGTC
1080

TGCCACAAAA GCCATGCTCT GAGGTCCTAT TCAGGAAAAT ATTAAGTCTG AATTATGGCC
1140

TCCCTTTCTT TGGCAGGGGG CACTGAGCTT CACGGAGACC CTAGAGTAGG AATCAACCCT
1200

TCAGGTGGAA GGCCCAAGGG AATTAGAGCA CTTAGAGCAG GAGACATGGG AAAAGAAAGA
1260

GGAAAAGAAG TTGGTAGGTC CGAAGGAGGC TGGAAAGAGG GGAAGGCAAA AGAATACACA
1320

ACCCAGCCGA ATCATGGGAA GTGAGAGATG GCTTCCTGCA ATCTAGTCTT GCGGGTTCTT
1380

GCTCTAGTTT CGGAAACTTC CCCCAGAGAG TCCAGCGCCG CGCACCTGCC CCTTGGGGCC
1440

GGTCCCTGCG CAGGAGCCTG CGGCAGGCTG GGAATGCCGT TCTGGGAGAG CAGCCTCCCA
1500

GGCGGGGTGG TGGGGAGACT GCAGGGAGGA GTTTTGGGGA GTGCAAACGG AAAGACTACA
1560

TTTCCCAGGC CGCCACGCTT TCCAGCTGGA GTCCTAGGGC GCTGACTGCT CCCAGTTTC
1620

095442-4100

CGTAGGGAAG CGCTGGGCTA CCGCGGCTAT
1650

(2) INFORMATION FOR SEQ ID NO: 2:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 1079 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: double
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- (vi) ORIGINAL SOURCE:
 - (A) ORGANISM: Homo sapiens
- (ix) FEATURE:
 - (A) NAME/KEY: exon 1
 - (B) LOCATION: 1..634
- (ix) FEATURE:
 - (A) NAME/KEY: exon 2
 - (B) LOCATION: 778..935
- (ix) FEATURE:
 - (A) NAME/KEY: oligonucleotide
 - (B) LOCATION: 727..746
- (ix) FEATURE:
 - (A) NAME/KEY: oligonucleotide
 - (B) LOCATION: 958..977

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

TGTGCGTCGC GCTCTCGCCC TCCTCTTCCC GCTGCAGTGT CTATGGAGCG AGGCTACGTT
60

TCATTGCCGC CCTGGCTTAA CCCTTCCGGC GCCTAAAAGG ACGGCCGGCC GGCCGGTCCC
120

TTGCACCAGG AAGAAGTCTT AGCAGCCAGC GGGCCCTGGT CAGGAAACTC TAAGGTACAA
180

GGAAAACAGT TGAGGAAGGA GCCAGAGCGC TCCGGTTTGG TCCTCGGGCT TCGCTGGGGC
240

GGGGCGCAGG CGTTGGCTTT AAGAAAGGGG AGGGGACAGT GCAATCCGGG TTGCCCCGCG
300

ATTCGGCCAA GGAATCTTCC GCTCGCTCCG GAGCGAGGAG CCTGTAAAGA GGCTGTTCCC
360

AGCTCCAGCT CTAACCTCGC CTACACCTTG GGCGGGCCCA ATGTCACGTT TGCAATTGCT
420

CAGGAAGGAT CCGGCCCGTC TCCGGAGGCA AGTCGGGCTG CGGTTTTTGC TGCTTATCTG
480

00747441300

GGAAGGCGAT GCCTAAGGGA CATGCTGCTT GCTAGGCAGC ACCCTGCCGG GATCCGACTG
540

CGATAGTTAG CTCTCCCTGG CCCTGAAGCC ATCGCCGGGG CGCCTGTTCT CTGTCCGGAC
600

CAGCCAGCGC TCCTCAGGAG TCTCACTGAA ACAGGTACCT GTCCTCCAA GGGACGGAGG
660

CTATGAGCTT CCTTAAGCGG GTCGCGCGCT CAGTCCGTCC CCTCTACTTC CTCTACTGTG
720

CCATTGATGC TCTCGGTCTT TGTGTCTTTC CCCTTTCCCC CTACTCCCGG CCATCAGAAC
780

CATGGGTCAT CCCCCGCTGG AGTTCAGCGA CTGCTACCTG GACAGCCCCG ATTTCCGCGA
840

GAGGCTCAAG TGTTATGAGC AGGAACTGGA GAGGACCAAC AAATTCATCA AAGACGTAAT
900

CAAAGACGGC AACGCGCTTA TCAGCGCTAT GAGAAGTAAG TGCAAGGCTT CGATGAGCTG
960

TTTCTCTGAN CTGGTGTGTC TGGCCTTTAA GCCTTTCCAC ACCACCAGGG GAAGGGAGAT
1020

TGCAGGGTGA CTCCCAGCCC ANATCTCTGA GGCAAATGGG TTTCCACAC TTGGGGAGT
1079

(2) INFORMATION FOR SEQ ID NO: 3:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 900 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Homo sapiens

(ix) FEATURE:

- (A) NAME/KEY: exon 3
- (B) LOCATION: 403..498

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 375..394

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 504..523

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

CTTTGAACTC CACACGTATA AAATGGAAAG CATCTAGTGT ATTGCCATA ATAGGGGTTT
60

09541423.42400

AGGAAATGTG TAAGGAGACC CTCTTAAGAG TATATCTAAG TAGTTTCATG TCTTCCATGC
120

TTTGAGTGGA ACAGGTCAGA GAGGAGAGGT GTTGAGGATA AAAACATGTC CCGTAACTTT
180

TAAGGACTTT ACTGAGATGC CCCATCCTTT CTTCTTGTA TTTAGTTGT TCAAACCTTC
240

TTCTACTGCA TATCTAATCT TTTGTTTTAT TTCATTAAAT GCTAGTTGCA ACCTGCTCAG
300

AGCACTTACC TTTGGTTTTTC TTTTATCTGC TCTTATAGAG ATGAGGAAAT AGATCAAAGC
360

ATAGCATTGT TGAACATTTT TCTTGTTACC CTCCTTTGC AGATTATTCT TCTGCTGTTT
420

AGAAATTTTC CCAGACGCTG CAGTCATTTT AGTTTGATTT CATTTGGAGAC ACTCTGACTG
480

ATGATGAAAT TAACATCGGT AAGTCTTCAG CTACATGTGG TCATATACCT GTTGAGGCAG
540

CCCTGAGACC ATGTAGTCTT TTTGATTGT GGATACAGAG CACTTGGACA TCTTCATCCA
600

CTGTGGTCCA ATGCCAAGGC CCTGGGAGGT TGATTAGGAA GGATCAGGAA ACTTTCCCTG
660

CCAGTCCCAT TTCCTCCTCA CACGACAGCA ATCAAAAGAT ACCCTTAAAC TTCTACTGAG
720

ATTTTTGACT CAGACAGTCT GCAAGCGACC TTTTCTTTAA AGCATAGTTA TTTTCCTAAA
780

GGATATATTA AAAGGGGGAC TTTTTGGTT TATTTCCCAA AATGGTTGGA GTTAGATTCT
840

TCTAAGGAAT CAAATTTCCC TAGAAAGTGT TAAATTAGCA TTTGTGTGTC TACAACCTAT
900

(2) INFORMATION FOR SEQ ID NO: 4:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 960 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Homo sapiens

(ix) FEATURE:

- (A) NAME/KEY: exon 4
- (B) LOCATION: 483..544

09581422.112100

(ix) FEATURE:

(A) NAME/KEY: oligonucleotide
(B) LOCATION: 418..437

(ix) FEATURE:

(A) NAME/KEY: oligonucleotide
(B) LOCATION: 551..570

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

TAGAAGAGTA GAAAGTTTGG GAGGGTGGAA AGGGTAATGT TTATTATTTA CTTTGTGTTG
60

ACACTACATT CACCATCCAA TTTAGTTTTT ACAGCAATAG TCTCTGGTTG GTCAGTAGTG
120

GAACAGGGAT CCAAACCTCGT TTCAGGGTTC TATAGCTGCC AATTAATTAC ACAGCAAACC
180

TCTTGCCCCCT TCCCTAATCT TTTAGCTTGT CTATGAAATA AGAGAAATTC TAGCTACTCC
240

TGAGTGGCTG TAAGGATTAA ATAAAATATT AAAGTGTATG GGGATTGATA AAAGAGGAAA
300

GAAAAGAAAA GAAACATTCA ACAGGTGCTG AACACCTGCT TTTGTCCTCC GATTTGACAC
360

CTTCCTCTAG TGGCCATGTG GGCATAGGGC ACTGGTCCCT ACTTCCTGTT GCACAGATCT
420

CTATCCATTT GTCTGTCAAG CACCATGATT AATTTGTTTT ACATTGATT CTCTCCTTCC
480

AGCTGAATCC TTCAAGGAAT TTGCTGAATT GCTCAACGAG GTAGAAAATG AGAGGATGAT
540

GATGGTAAGT CACTAACGCT GTCACTGAAG CTGAGTTCAT GGGTGATATA GGGGATTTTT
600

CCTTTCCTTT ATGCTTGGAT TGATCCTATA CTATTTTGAT TTCTGTCAGA TAGCTTCTTG
660

GTGCTATAAA AATAGTTAGG TAATAGATCT GGTATTATG TCTCAAGCTT CCACCCTGAG
720

AGTTTGGCAT TAGATAGAGG GAAATTAACG TGCAAATCCC ATCTGTGTTT ATTTCACTGA
780

AAAATAATTT CAGTGGATAT TAAACTGGGC CTTTGAACAT GTTGACAGAA ATTGAGGTCT
840

TTAGTGTTTT TAGCCAAATT ATCCATTTGT TAATCTTTAA TTTGTGGAGT AGTTTTACTT
900

TTATAGAGAA AATCAGTAGA AAATAAAGAT AGAACTCATA TACCACCTTT CTCTCTCCCA
960

09541433 112400

(2) INFORMATION FOR SEQ ID NO: 5:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 960 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Homo sapiens

(ix) FEATURE:

- (A) NAME/KEY: exon 5
- (B) LOCATION: 451..522

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 423..445

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 553..574

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

TCTGGTTTTA TGGCACAGGC AGGACTGTAA ATGTTTTATC CTAATGCTAT CACTGCACAC
60

TCATTGCCT CTGATGAAAT GATTTCACCT TGCTGCAATT GTCCTTTTCT TGTATTTGCC
120

ATTCTCTTTC TCTTCTCTTT TTCACCTGTC CTCAGGTCT CCTGTGTCCT CTGCTGGACA
180

TCACTTCAGA TATTTATCGA TTAAAACTC AGGTCAGACT ATCATTAAAG TTACAGAGAA
240

ATGCCCTCT TATTCTTTCT CCCATTCTT CTCAATGCAT TTGATTTTTC AGAAACAATA
300

TAGAAACAAA CAGTAACAAA ACCCAACAAA TCAGCAAACC ATTTAACATT TTGCAGGTTG
360

GTATATAAAT GAAAATGTAG TAACAAGGAA TCTTGTATCT GAACCTTGTT AACCTAGAAA
420

TTGTTTTGTT TGTTTTTCCT TTTTGTCTAG GTACACAATG CTAGTGATT TCTGATTAAA
480

CCCTTGAAA ATTTCCGGAA GGAACAAATA GGCTTCACCA AGGTACATTT TCTGTATATG
540

CATAAGATTT TTAAAAATAG CAATCGAATA GTTGTATGGG CTAATATTCT TCACTTTACA
600

09584430 413100

AAGATATGCA CCAATCTGCT GGTGCTTTGC TCTTGGCCTA GTCAGCCTCC TAAACTGTGC
660

AAAATAAATG TTTGTTGTTT ATGTCACCCT GTCTATGGCA TTCTGTTATA GTAGCCTCAG
720

CTAACATGAC AAAGGGGGTG GGGAGGTGGG TGATTAGTTT CTATGAGAAA ATGATCACGA
780

AAGAGAGTAA GAAAATCTAG AATTGGCCTC TGACTTTGTG GCCAACAGGC TCTGTATCTG
840

TGCATAAGTT TCTTCTTCTT TTGGTGTTTT TGTTGTTTTG TCTGGAAAAC TAGCTAGCTA
900

TCATGTATCA ACTGCCTGCT ATATTGAGCA CTAGGCTAGG TGCTTTACGT TCATTCTTTA
960

(2) INFORMATION FOR SEQ ID NO: 6:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1020 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Homo sapiens

(ix) FEATURE:

- (A) NAME/KEY: exon 6
- (B) LOCATION: 416..517

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 388..407

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 540..559

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

CATTGTAGAT TAACCTTTTC ATGACAAATA TTTACTTCCA TTTGTGTTAT GTCTTCTTCC
60

TTTGTAAGAA AATTTGAATG AGTGGTTAAT CTATGTGAAA AATATTGGAG GGAAGAAAAT
120

ATATCTACTG CACAGGCCCT TTTAAGGTAT CATTCTCTAA GGAGCAGCTT CCATAGCTTT
180

CAGCTGTAAA AATAGGGACT GCCATTTCTG CAGGCAGAAT GGTTTGGGGT TATATTTTCAG
240

00221 224T6660

GAAGCTGAAA CTGCTGAGAC CAATACAAAG TTAATTCTCC GTTGCTTTTT TTTCTTCCA
300

GGAGCTAGAG GGATTGACCA CCTGAAACCT GACACTATCT CCTTCATTCT CCTTTCTAAG
360

CAGACAGAGT TACAACTACA GACATTTAAT CTTTGCCCTT CTTTCCCACC TTTAGGAGCG
420

GAAAAAGAAA TTTGAAAAGG ATGGTGAGAG GTTTTATTCT TTAAGGATC GGCACCTACA
480

CCTGTCTTCA AAAAAGAAAG AATCTCAGTT ACAAGAGGTA TGTTACACAA GCCTGCCCCT
540

GCCTTCCATT GCTAGCTATG CCTTAGAAAC AGTGTGAATT TTGTAAGCA AGGCTTTTCC
600

CATACCCCGT CTCAGCAGGG AACCTCATGT GATAGTAGCA CTTGTAGTCA AACTGTGGC
660

CTGAGACTCA GAAGCCCTGA ATACTAAGCC AGCTCTTCCA CTAAGTCAGG GTGTGACCTT
720

GGATAGGATA CTTCTTTCTT TGCCTCATTG CATTATCTGT AAGAAGAGGA GTCGAGAGTC
780

CCTTTCAATT GCAAGTCCAA AATCCATGCG AGGATAAAGT TAAACTAGTG TTGTATTGG
840

TAGAAATCAG GAACAGATCT TCTACTTTTT TCCCTGAGAG ATTCCACAAC CTTTTTTTTT
900

TTTTTTTTTT TTGGTGAGGG GTTAGGGGAA TGTTCTGTTT TGGGACTACT GGTACCTGG
960

GAATTGCAGT GCCTTCAGTT CAAATAAGCT ACCATTCGGT GAGACCCTAC CACATGCCAT
1020

(2) INFORMATION FOR SEQ ID NO: 7:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1020 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Homo sapiens

(ix) FEATURE:

- (A) NAME/KEY: exon 7
- (B) LOCATION: 464..574

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 436..458

09581423.12400

(ix) FEATURE:

(A) NAME/KEY: oligonucleotide

(B) LOCATION: 584..603

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:

CATTTCTTGT TAAATAAGAC TTGAGACTTG CAGTTCTTTC TGTGTGAATC CCCTCAGTAT
60

TGGCCATAGT AGCATGCTTC TGTTCCTGGG CTTTGGCCCT TGTTAATTTT TTATTCTCTG
120

TATTCATGTC TCTGTAGTGT TTGAGGGAGC AGATTTTCCT CTGACTTCAG TTCTCTGTTG
180

AATATAATAA AATGTGTGGA TTTTACACTT GTTCAGCTTT TTTCTTTTTG TCAGGGTGGA
240

TGTAATGACT TCAAGCTTTT TATGTGTTGG ACCAGAACT GGAATCCTGT ATCAGTCACT
300

TTTTTATCTC TTGGCTTTAG GTTCTCTATT TCTAAAATAA AAGGTAACAT CACTAGTTGG
360

TGGCTAAATG CTCATCTAGC TCTAGCATTG TGTTCAATCC TTAAGTACTG ACTTTCATGA
420

ATGAATATCC CAATATGTAA TGTTTGT TTTT CCTTCTTCTA CAGGCAGACC TACAGGTGGA
480

CAAGGAGAGG CACAATTTTT TCGAGTCCTC TCTTGATTAT GTTTATCAAA TCCAGGAAGT
540

TCAGGAGTCC AAGAAGTTCA ATATTGTGGA GCCTGTAAGT TTTCTCTGTT GATGAATGGT
600

CTAAAAATAT TTATCAAATG CCTGGTAAAT GTACAAACTT TGATCATAAA AACAGAATCC
660

CAAAGAACGT GAAGGATAAT CAGTGGAAT GTCTAGAACG TAGTCTAACT TACGTTTTTC
720

TTGTTCTTTA GTATACTTTT TTTCTCTCTT TCCCCCAAC CCTTTTCTCT CTTTTTCTAT
780

GGCTATTTTC TTTCTCTTTC CTTTTTTTCC TTTCTCTTTC CTCCCCCTT CCGGAATTTT
840

TCTACAACTC TTCCTGTTAC TTTGTTTCCC TCCCTTCCTT CTTCTTGCGC TTTCAGAAGC
900

TTCTGAAAGT CTTTCTCATA TACCAGATAC TATGCTAGGG ACTCATGTCT CTTGCAGTCA
960

GTGACTTCTC ATTCTACTAC TTTTACTTCT GTCTTGTGTA GTTTTTTCCC ATTTACAGTT
1020

09581422 12100

(i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 479 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: double
 (D) TOPOLOGY: linear

(vi) ORIGINAL SOURCE:
(A) ORGANISM: Homo sapiens

```
(ix) FEATURE:
      (A) NAME/KEY: oligonucleotide
      (B) LOCATION: 219..239
```

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 8:

CTTGGAAATTC TAGGTTACTG CATCCTAGCC AATTTGTTAG AATATACTGG TGTGGATCA
479

(i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 600 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: double
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

(A) ORGANISM: Homo sapiens

(ix) FEATURE:

(A) NAME/KEY: exon 9

(B) LOCATION: 134..263

(ix) FEATURE:

(A) NAME/KEY: oligonucleotide

(B) LOCATION: 108..128

(ix) FEATURE:

(A) NAME/KEY: oligonucleotide

(B) LOCATION: 336..355

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 9:

ACCATTTCTT TTTTGTCTT GTGGTGCAC ATGCTGTAAG CAGGGAAAAC TTTGTACTGA
60

GTCTCTGACC AAGAAATACT TTTTCATGAT AATGATGATG ATAATAATGA TTTTCATGAT
120

GATGTCTTGA CAGACAAGAA ATCATTCTC CAGTACCCGG GAAGAGATGG AAGAACTTAA
180

GAAAAGGATG AAAGAAGCTC CCCAGACATG CAAACTTCCA GGACAGCCAA CTATTGAAGG
240

CTATCTCTAT ACACAAGAGA AATGTGTGTG GGGACATAGG GGTATCCATT GGGTTTCAAT
300

AAGCCAGGAA GTACTGCCAC TTGTCGGCTG TGAATTTTGG GCACCCCTTAC TGTTCATAGA
360

CCCCTGATAG CTAAAATTCC CTTGGAACGC AGGCAGGGAA TACTGAAAAC AAAAAAAAAA
420

AAAGGAGAAA CTGAGAGGAA GTTAAAGATT TGTCTTACAA AGGCTGTGTA GTGATAAGAC
480

CTAAGGTTTT CTCTGAGATT CAAAATGGGT ATTATTTGTT CTTTAATCCT TCTGATTATT
540

CTTTTGAAA AAAGGGAAGT AGAGGAAAGG AAGTAGAAAA ATAATATTTT TTATACTTAT
600

(2) INFORMATION FOR SEQ ID NO: 10:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 1020 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: double

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

09581422 1200
001211 22472560

(A) ORGANISM: Homo sapiens

(A) NAME/KEY: exon 10

(B) LOCATION: 383..483

(A) NAME/KEY: oligonucleotide

(B) LOCATION: 361..380

(A) NAME/KEY: oligonucleotide

(B) LOCATION: 492..511

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 10:

AGAGAGATGT GTAGAGTCAT GTACAGGTTA GAAGGGGCAT AACCCAATGA CCAATGACAT
60

GGAACAGTTA TAGAGACAGC AAAGTAAATG ACAATAGTCT CCAGCTTCTG GGATATATGG
120

TGATACAATT TATAGGGAAC ATGGCAGGGA GAGTAGGTTT TGAGCAAGCC AGAAGCAGCT
180

GTGGGAAGCA GTTGGTGAAG GTTAGAATCT CCCTGTGTGC AGTAGGTAGG TGGCTATGGA
240

AGGAGGGCGG TCAGGGCAAG GGCAGGGCTG GATCTGAAGT TTGACTCTGA AGAGCAATGT
300

GTAAATAGCT TCCATCTTAG GGTGACTTC CTATACAGCT AAAATAGTTA TTCTGTCTGC
360

TCACCTTTTAC TTGTCCTTGT AGGGGCTTTA GGAATATCCT GGGTGAAATA CTATTGCCAG
420

TATGAGAAAG AGACCAAAAC ACTGACCATG ACGCCTATGG AGCAGAAGCC AGGTGCTAAG
480

CAGGTCAGTT CTTGTTTGCA CCATATTTTT GGAAATGGAT CTATGACTGT TTCTCAGAAA
540

AGAATATATG TTGACCTAGT ATCAAATCAT CAAGTTCATC ACTGTTACGT GAGGCCATGA
600

CTTTATATGT ACACCTTGGC CTAAGTTTGA GTCAGATAGC ACTGAGTTGA GTGAAAATT
660

TCTCTGTTGA TTAGAGCAAG CCTTTTGAAA GTGCCGGTAG TCTTTCAAAC CAGTTATTTT
720

TACAAGTGCC AGTCACATTG TACAGTCAAC TATGTAAAAA TATGGATGAA TTACTTTTAA
780

GAATGCTCTA CTCTTGGATT CTTTAAATA GCAAGTTTTA AAAATATGAA TTGAATTCCA
840

[illegible]

AAATTCCTTT TTTACAGGAG TGTGTTTATG GCCCACAGTT GGAATAACCG ATACTCACAT
900

TCTATGTACT ACTCAAATAT CTTTAAGCAG TTAATCTCTC TTTTCTGCC CTCCAAACCT
960

TCTCTCACTC CTGAAAATGA CAAGATAAAT TTAACACACT GAAAAAATA GTTACTTAC
1020

(2) INFORMATION FOR SEQ ID NO: 11:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 294 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Homo sapiens

(ix) FEATURE:

- (A) NAME/KEY: exon 11
- (B) LOCATION: 107..198

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 81..100

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 223..242

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 11:

CCCTAGACCA GACGCCTTCC AAGTGACACT GAAGACACTG TAGGAAGCAG AACAGCTTC
60

TGTGTCTTTT TTTAAAAATT GCCTGTTAAT TCTTGTTTCA TATCAGGGGC CCTTGGAATT
120

AACACTGAAG TACTGTGTGA GAAGGAAGAC GGAGTCTATC GACAAGAGGT TCTGTTTTGA
180

CATAGAACT AATGAAAGGT AAGCTGTGCC GCTGTGAATT GGCAATGTCC CCACGTGCCA
240

GATGCTTAGC CTGGGTATGT CTTTTATTTT CCTCCGTCAT CCCACGTTGA TGAC
294

(2) INFORMATION FOR SEQ ID NO: 12:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 494 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double

09581433.112400

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

(A) ORGANISM: Homo sapiens

(ix) FEATURE:

(A) NAME/KEY: exon 12

(B) LOCATION: 211..289

(ix) FEATURE:

(A) NAME/KEY: oligonucleotide

(B) LOCATION: 188..207

(ix) FEATURE:

(A) NAME/KEY: oligonucleotide

(B) LOCATION: 300..319

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 12:

CCACCCCGC CTCCATGGTT CCCAAAGTTA CTGTTCTGTA AATTGCCTAT TGTTCCTTCTT
60

GAGTAAAAGA GATTTTGAAA AATTAGTATT CTGTAAGCCT ATGCTTGCTC AGTGTGACAA
120

ACTCCAGGGC AGAGAGGGAC CCTAGATCAC AAGACTCCAT TCTCTCAGTT GAATTTTCTG
180

CTTTATACTT ACCATTTTTT TCCCCCTCAG GCCAGGAACC ATCACTCTGC AGGCCCTTTC
240

AGAAGCTAAC AGAAGGCTAT GGATGGAAGC CATGGATGGG AAAGAACCTG TAAGTTACCT
300

GAACTGGGG CAAACCTCCC CAGCATATGC CAGTGTATGA GTGCCCTCTA GTGGTATCAG
360

TGGGTCTCAN ACAATTAAAT GGTAATGGAT TGTTTAGTCT CAGTTTCTAG GCTGTAAGGA
420

ATTGTTTCCA CATCTCTTAG CAGGTAAGGC AACTGGAGTT CCAGAAAGGT TGAGGGACTT
480

TTCTGAGACC ACCC
494

(2) INFORMATION FOR SEQ ID NO: 13:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 378 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: double

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

09584427.12100

(ix) FEATURE:

(B) LOCATION: 212..245

(ix) FEATURE:

(A) NAME/KEY: oligonucleotide

(B) LOCATION: 166..189

(ix) FEATURE:

(A) NAME/KEY: oligonucleotide

(B) LOCATION: 259..278

(xi) SEQUENCE DESCRIPTION: SEO ID NO: 13:

GGCATGAGCC ACCACGCCCTG GCCTGTTCAA GTATTTTCTA GCAATCTTGG CAAAGCAATT
60

ATGTTTAGCC CACTTGGCTA TCTTTTAAAC ATCCTGGAGT TTCTAATCAT TTTTAATGCC
120

TATCTGGGGA AAGATATTTA ATATTATGTT CTCTGTTTTT CTATATTGAT TGACAATAGC
180

CATGGATCTT TCTGTTTATC TTCTTTTGTA GATCTACCAC AGCCCTATAA CAAAACAGCA
240

AGAAAGTGAG TCACTTAAGT TTTTGGTCTA CTAGCATTAT AAACTGCCAG CTGTCCGATT
300

CATAGTAAAT ACCATCATTA ATGATGTGTA CTACTAACGC AAGTCTGAAT ATGGATGCCT
360

TTGTGTGAAA TAAAATTC
378

(2) INFORMATION FOR SEQ ID NO: 14:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 393 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: double

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

(A) ORGANISM: Homo sapiens

(ix) FEATURE:

(A) NAME/KEY: exon 14

(B) LOCATION: 172..234

(ix) FEATURE:

(A) NAME/KEY: oligonucleotide

(B) LOCATION: 133..152

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
(B) LOCATION: 250..269

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 14:

AAAATGTAA ACCTCCCCTG AAAAAATGAC TCTTCCATT TAAGGGTGAC TAGAAATGAG
60

CAACTAAAA CCCTTAGCTC TCTCAATGCA GTCCCTTTGC ATGGTCATTA AATGTTTAAT
120

AGGTGACACC TGTTCAGCA GGATCTAACT CTTTTCCTTT GCTTGAAACA GTGGAGCTAA
180

ATGAAGTGGG CTTCAAGTTT GTCAGGAAGT GCATCAATAT TATTGAGACC AAAGGTAAGA
240

TCTGAACCAT AGTCTTGGCA TTGTCTGAAT CTCGTCCTC TGATTTTATC CTGGGCAATT
300

TCTCTGAAGT AGCGTTTTAG GAATGAAGAC TGTTTATAAA GCTTGTGTAG TAGATGCAAG
360

CTAGAAAATT TCAGAAAATT CTAAACTAGT GGT
393

(2) INFORMATION FOR SEQ ID NO: 15:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 436 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Homo sapiens

(ix) FEATURE:

- (A) NAME/KEY: exon 15
(B) LOCATION: 207..281

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
(B) LOCATION: 151..170

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
(B) LOCATION: 293..315

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 15:

AGATTAATGA GGGTTTGGTA CACTCCAAAT GAAAGGATGG ATAATTTGGA GAGATGCTGT
60

AGGACTATTC CCCTGTTACA GGAAGGCTG GAGAACTTGG AGTATGTAGT GTGACCCCTT
120

09581422.12100

CCTATCTGAA TTGACTCTAG TGTACCAAGG GGAGATGACA ACTTTAGCTA TACAAGTGAA
180

ATTAACCTGA TTTTTCCTC CACTAGGGAT CAAGACAGAA GGGTTGTACC GCACTGTGGG
240

CAGCAATATT CAGGTCAGA AGCTGCTGAA TGCCTTTTTT GGTAACAATT TCACTTTGAT
300

AATTCCTTATT GGGAGTACTT TATGTGTTAC AAAGAAATGT GACTGGAAGA GAAAGGAGAC
360

ACTGCTAAAA TGTGGTAGAA TAGTTGAAAA AAGTATTTTC TAAAGTAAAA CATACACATA
420

CTTGCCCACC CTGGGC
436

(2) INFORMATION FOR SEQ ID NO: 16:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 547 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Homo sapiens

(ix) FEATURE:

- (A) NAME/KEY: exon 16
- (B) LOCATION: 270..354

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 221..244

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 363..382

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 16:

ACCCAGTNA TNTGATGAAT CTAAGAAGAG TNGATNTTGT TTGTTCAAGT TTTTCTTGTT
60

GTGTGAATTG GATAGATTAC TTTNTTATTT CTTATATGGC AGACCAGAAT GCAGTCATGT
120

TTTTGAAATA TCAAAGATTT GCTTCTTCTA AAGTTTTGAT NTCTTAAAAA CTACTTAGGG
180

TNATATACTT TGTTTTTCTT TTAAAAGAGG GAAAATGTAA GATTTTTTTG ATGATTAAC
240

095442.412400

TTTGTTTTTT GTTTACTTTT CTCAAATAGA TCCTAAATGC CCAGGAGATG TTGATTTTCA
300

TAATAGTGAC TGGGACATTA AGACAATCAC CAGCTCCTTG AAATTCTACC TCAGGTATGC
360

CTGATTTGAA TTGGGAGTTT GCTTTTCATA GCTGGTGAAA TTTCTCTGGG TGTGAGCGG
420

AGTTAACGTG GTCTCAGTTC CAGGAGTTTG GATACAATTG CTTAANAAAA AACATGTGAA
480

GAGGATTTCT GGCCANGAAT GTGCAAANAC TGTTTTTTTAA ATCTGAGAGT TTAAGCAAGA
540

GAAGCAT
547

(2) INFORMATION FOR SEQ ID NO: 17:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 601 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Homo sapiens

(ix) FEATURE:

- (A) NAME/KEY: exon 17
- (B) LOCATION: 355..413

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 305..324

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 438..457

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 17:

GGACAATTGC AAAAGCACTT CGGAAATTCT AAGGATCTAT CAAATCGTAA GGGATTCATG
60

GTAGCATTCA GCATGGGTCC CCTCTGGAAT TTTGCAGGAC TGGTTTGTGC CTCTTTTAC
120

TTNTGGGAGC TAGTTGGAGA CCTTGCTAGA GGGCTCAGCC CATGCTTTTG CAGGTCTTTT
180

GTTGAATTAC TAGCAACTTG GATTCCCTGA CGAAGCTTCA GGTGAAGAGA AAAATGTATA
240

TAATCCCACT AAGCTGTAGG GCTCAGGAAC TTCAGCCTTG CTGTCCCCAG AACTAAGAAT
300

005442-1100
004211-22443560

CCAATACCCA GCTGCTTTNT TCCCAAAGCA ACTGACAATT TTCATTCATT TCAGGAATCT
360

TTCTGAACCT GTCATGACCT ATAGACTTCA CAAAGAGCTG GTCTCTGCTG CCAGTAAGTA
420

TTTATGTTAC TAATTAAGTG TGTGTCCTA GTTCTTAAT GTTACTGCA ATAAGCCTAG
480

AAAATTGTTT GAGGGGAAGT GATTGAGGGC ACAGAAACCT AAAACACATA CACAAATTAT
540

GCACAAGTGC CAAATGAAAG TATTCTTGCT TGCTGTCTAA CTCAANAATT CTATTATTTT
600

T
601

(2) INFORMATION FOR SEQ ID NO: 18:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 387 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Homo sapiens

(ix) FEATURE:

- (A) NAME/KEY: exon 18
- (B) LOCATION: 80..185

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 25..44

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 218..237

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 18:

GTGCATATAT ATGTGAGAAT TTTGCTCAAT CCAGTAGCCC AGAAAGCCAA ACCATTTATC
60

TCTTACTGTT CTATCCCAGA GTCTGACAAC CTGGATTACC GCCTAGGAGC TATTCACTCC
120

CTGGTATATA AGCTACCAGA AAAGAACCGA GAGATGCTGG AACTTCTGAT AAGACACTTG
180

GTCAAGTAAG TAACTGCTGG ATTTTCAGAA AAAGTTCCTA TTAGAGGACT GGCCCATGTG
240

095442 112400

GAGTCATCAT GCATGTGGAT TGTCTCA
387

TAAGCTTATA ACTATATTAT TTTCAGACAC CATGCTAAAT
460

[illegible]

(2) INFORMATION FOR SEQ ID NO: 20:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 512 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Homo sapiens

(ix) FEATURE:

- (A) NAME/KEY: exon 20
- (B) LOCATION: 230..377

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 206..225

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 383..402

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 20:

TCTTCCAAAC TGAGAAGTGC CAGGTTNTGT GCCTTGAGCA TAGTAGGAGN TACNTAAACA
60

TTTACCTGTA GNTAGAGTGA TTAAGAAAAT CTCTGATTCT TTGAGTCATG TTAGTATTCA
120

CGTNACAAAC TCTAGATATA AGGCCAACAA GCATCAANTG GTGGGTAGCA TTCAGAAGAC
180

AAAAANTTGA TNTAANTATT CTNTAGATAT NTTCTTCTT TNTCCACAGA TCTATTTAGG
240

TCCACCTGAG GAAAGCGCTG CACCGCCAGT GCCTCCGCCT CGGGTGACAG CAAGAAGGCA
300

CAAACCAATC ACGATTTCAA AGCGCTTGCT GCGAGAAAGG ACGGTTTTCT ATACTTCTTC
360

CCTGGATGAA AGCGAAGGTC AGTACTNAGG TTCTCCTTTA GCTTCTGAAT GGTGATTAGA
420

CACNNAGNAN GATATCNAAT GGCTCAAGCG GTGGCATCAC CATTTNTCTC TCTATAAAAG
480

TANACCTTTC CTGNCTCCTG AACTTAAAAG CA
512

(2) INFORMATION FOR SEQ ID NO: 21:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 841 base pairs

09504427.112400

(ii) MOLECULE TYPE: DNA (genomic)

(A) ORGANISM: Homo sapiens

(A) NAME/KEY: exon 21

(B) LOCATION: 185..508

(A) NAME/KEY: oligonucleotide

(B) LOCATION: 151..170

(A) NAME/KEY: oligonucleotide

(B) LOCATION: 511..530

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 21:

GGCTTTTACAT GAACTGGGAA GGGTAGAGAA TGATTTTGTG GGATATAGTT GGTTCGTGCC
60

ACAGTGACAT AACTGCTTTG AAAATGTATA CAAATTTTCA AAATTAAGTA TGTATGCATG
120

TATCAAAATG AAAAGGTTTT AAAAGTTATC ATTAATCTTC CCTNTTGGCA CCAACTTTTC
180

CTAGATGAAA TCCAACATCA AACACCGAAT GGTACTATCA CCAGCAGCAT AGAACCCCCC
240

AAGCCACCAC AACACCCCAA ACTACCTATT CAGAGGAGTG GGGAAACTGA TCCTGGGAGG
300

AAGTCCCCAA GCAGGCCTAT TTTGGATGGC AAGTTGGAGC CCTGCCCAGA GGTGGACGTG
360

GGGAAGTTGG TGTCTAGGCT GCAGGATGGA GGGACCAAGA TCACCCCAA GGCCACCAAT
420

GGACCCATGC CAGGCTCTGG GCCACCAAG ACCCCCTCTT TCCACATAAA GAGACCAGCT
480

CCCCGGCCCC TGGCCACCA CAAGGAGGGT AAGTGCTTGG GAATCCCATG GGAGCCAGAG
540

CTGACCCTAA CTACTTTTCA CCTTGAGATC CTTCTGAGTT TGGAGATATA TTTAAGTGGA
600

660 AATAATGTTCC AGTTTACTAC CACTAATATT GGAACAGTGG GCAAGATCAC AATAATCAGT

CACAATAATC ACTAGAAATGT AAGCTCCATG AGGGCCGGGA TTTTTCACCT GTTTTGTTGA
720

```

# 1. Importing the libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')

# 2. Reading the dataset
data = pd.read_csv('data.csv')

# 3. Data Preprocessing
# 3.1 Handling missing values
data.isnull().sum()

# 3.2 Handling categorical variables
data['category'].value_counts()

# 3.3 Handling numerical variables
data['age'].describe()

# 4. Exploratory Data Analysis (EDA)
# 4.1 Univariate Analysis
data['age'].hist()

# 4.2 Bivariate Analysis
sns.pairplot(data)

# 5. Feature Engineering
# 5.1 Creating new features
data['new_feature'] = data['age'] * data['weight']

# 6. Model Building
# 6.1 Splitting the data
train_data, test_data = data[:1000], data[1000:]

# 6.2 Training the model
model = LinearRegression()
model.fit(train_data[['feature1', 'feature2']], train_data['target'])

# 6.3 Evaluating the model
train_data[['feature1', 'feature2']], train_data['target']

```

ATCCTCACAA TCCCACTGTA AAGAATGTAT TATTCTTAAT ATTTTCTTTT TTTTTTTTTT
840

T
841

(2) INFORMATION FOR SEQ ID NO: 22:

- ```
(i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 615 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: double
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:
 (A) ORGANISM: Homo sapiens

(ix) FEATURE:
 (A) NAME/KEY: exon 22
 (B) LOCATION: 320..485

(ix) FEATURE:
 (A) NAME/KEY: oligonucleotide
 (B) LOCATION: 294..313

(ix) FEATURE:
 (A) NAME/KEY: oligonucleotide
 (B) LOCATION: 496..515
```

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 22:

TTTAATCCTC CCACTATCTC TGTAAGATAA TATATTGTGG ATCTTTATTA TATAGCTGGG  
60

GAAACTGAGA CTTAGGGAAT GGATATGACA CACCCAAGAT ATNTGAACT CCAGAGCTGG  
120

GGTTCAAATA TAGACTTTCT GAAGGGACAG TTGCCAGAAA AATTACAAAA AAAAAAAAAA  
180

ATAGCCAGAG TTGTTAGTCA CCAAGAAGAA ATGGAGGCCA AGGAAGTTGG CCCAGGTAAC  
240

TCTCATATTG GGTGCCTGCT CATGAGTAGT GTTCTGTTTG GCTAACCATC CAAGTTCCTG  
300

GTATCATTTT CTCTTCCAGG GGATGCTGAC AGTTTCAGCA AAGTGCGGCC TCCAGGAGAA  
360

AAGCCAACCA TCATCCGCCC CCCAGTGAGG CCCCAGATC CTCCCTGCCG GGCAGCTACT  
420

CCCCAAAAGC CAGAACCAAA GCCAGATATT GTGGCTGGCA ATGCGGGGGA AATCACATCA  
480



TCTGTGTCAG TAGGGTTGTA CCTCAAAGTT GACTGAAGTC CTGTACTAGG CCACTAGGAA  
540

TGCTTTTCAGG ATCACCATAT TAAGGGTATA CAGTGCACAG CCCTGGGGCA TCCTTCACTT  
600

TATAGTCTAG GGAAA  
615

(2) INFORMATION FOR SEQ ID NO: 23:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 475 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Homo sapiens

(ix) FEATURE:

- (A) NAME/KEY: exon 23
- (B) LOCATION: 211..261

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 179..198

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 271..291

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 23:

AATGGGGATA AAGAGTGTC GCTATGGCCT TAGGGGTGCC TATGGGCTCT GGGCCATTTC  
60

ACATTTGTAT GTGTAGGGCC TTGCCAGCA AAGGCAGGGG CTGGCATTGG TGTCCCATCT  
120

GGTTCAGAGT CTCCTGTCCT TTCTGTTGGC CATTGGTTCT CACGTGTATA CCAAAGCAAC  
180

TTATGGGACT TGGTTGGCTT CTGTTTGCAG GGTGGCTTCC AGGACCAGGT TTTTGAAC  
240

AGCTTCCCGG AAAACAGGAA GGTAAGATAT GGAGGTGACA AAAGAAAAAC CAAATCGCCT  
300

TTTAATAACT GCATCCTTAG CATACAATTG TGCTCACTCT AACATCTTTC TCTTTTGT  
360

TCTCTACAGC TCTGTCTCTG TCTGTCACTT TCTCTTCCCC AATTCTGTCT CTCCATCCCT  
420

09581433.11300

ATCTGTCTGT CACCTGTTCA CCTGTGTGTC TATTTGTTTC TCTCATATTC TTTT  
475

(2) INFORMATION FOR SEQ ID NO: 24:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 238 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Homo sapiens

(ix) FEATURE:

- (A) NAME/KEY: exon 24
- (B) LOCATION: 115..156

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 65..84

(ix) FEATURE:

- (A) NAME/KEY: oligonucleotide
- (B) LOCATION: 165..184

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 24:

CCATTTTTC AATGGCCTCT TTAGCACTGG CCTAGAAGTG TCCCCATTG CCCCAATTTA  
60

CCTTTCCAGT CCGATTCTCT AGAATCTTAG TGAAACGTCT TTCTTTATCC ACAGTTCTCA  
120

AGGCAGACTT CCGGAGATG AAAGTTGAGG CTACAGGTAT GCAGTCCCCA TCCCTGATTA  
180

CAAAATCTTG TTCCACATAA GCCTTCATTA CGGGATCTGA TATTTTGAGG ACTGGAAT  
238

(2) INFORMATION FOR SEQ ID NO: 25:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 4504 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Homo sapiens

(ix) FEATURE:

- (A) NAME/KEY: exon 25
- (B) LOCATION: 1..4235

09504429.11100

## (ix) FEATURE:

(A) NAME/KEY: polyA\_site

(B) LOCATION: 4236..4241

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO: 25:

GTTTTAAAAG CCTTGGCCTC AGAGGACCCT TTCCAGGTTC TGAAAGAGTC TTTCTTCTTC  
60

AAACCTTTGT GTGCGGAGTC ATTTTGTGTT GAAGAGCAGC TCCTTCCTAG CCTTGCACTT  
120

TCAGACTCTC TCTGGGAGGC CATAAAATAA GGAGCATATG TCCTAGACAG GTGTTTATAT  
180

CTCCTTTGTA TTCTGTCTTC ATCCCCTCAG AAGGTCTGTT TTGAGTTCCT ATAACACTGT  
240

GAAGAGCTGG ACTCCCTCAA GCCAGACTCT GCCAAAACCA AGATATCCAC TTACCTGAGT  
300

TGAAGAGGGG AGCTCAGTTT TCAACTCTTC CCTGAACTTC CTGCTTCCTC AGAGGGCCAT  
360

TGAACTCTGA GAGATTTGGG GCTAAAGACT GATCTCAGGG GTCTTACCTT GAACTGAAGG  
420

CCACTTGAGT TGGGGCCATT GCTTACCTTG GTTGAAGGG AATAGAAATG TTTGCTGAAC  
480

ATTGGAGAAT CTCAACATGT CTCCTACTGA GGATATGGAC ACTGGTGCCA TGTCAGCGCT  
540

CTGGTGCTGC AGTATGTTGC CAAGAGCCCG TCTGCTCCTG CGAGGCTATG AGTGGGATGA  
600

GTGATGCCCC CACAGCACCT CCATGTGGAC TTAGGAAGGT GGCCTTCCTG CTGTTACATG  
660

CAGCCACTTA GGACAAATCT GCAAAGCATG TTTTGCATGT AAAAGCCTAG GTCTATTTGG  
720

ATTATTCTTT CTCCTTTTTT TTGACAGCTT CCTGTCAAGC AATCAAGAAA CAAACAAAAG  
780

CTGAACACAT TTCCTTTTAA AAAAAGGAGA CTGTGTTTGG TCCTGTAGGA GTTCTATTTT  
840

GGGGTCAAAT GCTAGAAAAA TTGTTAAGGT GGATTGAGGC CAGGCAGCTG TCACTGCTGC  
900

TTCATGTTTG CCTTCTGCAC ATAAACTCTT TTATCTCCTG AAAAAAGCAG TTCTTAACCC  
960

AGTGTCCATG GACTCAGAAA CTCCATGATG CCCCTGAGAT GGTATGCACA ATTCCATGAC  
1020

09501422 112100

AATATGCCCT TTCTGGGGAG ATAGTCCATA ATGTTCTGCT AAATTTCAAA TGGGCTCGTG  
1080

ACCCAAAAAA GTCAAGAACC ACAGCACTTG AGTTAAAATA CTCTTTTAC AATCCATATA  
1140

AGCCCTTGAT TGGAAGGGCT TTTCAAATC ATTTAGTCTA ACAACTGCCC AGTTTCCAGT  
1200

CGGGGGAAGT GAGGCAGAGC AAGGTAGTGA TCACACCAGT ACAAGATTTC AGGTCCCAGG  
1260

CTCCTATGCA AGTTTTTTTT CCCCATATA TCACACTTAT TTAGCAAGGG ACCTTGTTGGT  
1320

TTGTGGCTTT AGTGGCCATC ATTTCTGGGG GTTGGCTTTT ACCCTTTTTC TTGAATATTT  
1380

GCCACCAAGT GAAAAATGTT AGGACATAAA CCCTTGCCAG GTCCCTTTCA TTTGCTATCT  
1440

CTATTATTGG AAAGGACCTA AAAATTGGTG TAATGGGGCA GAAATCTGAG GAATGGACAT  
1500

TTCTAATTCC TGTTTGTGA AGGGAAGTTG CTGGAAAGAG CATCAGTACT TGTTTCTATG  
1560

CAGATGCCTG GGCCGTAGCT TGTCTGTAGC GTCTGTATAA TTATAATGTT GCCCAGTGTG  
1620

AGGGAAAGAG CTTTCCTACT TGCACTCTTC TACCAAGGCC CTGTTAGTGC ACTGATTATA  
1680

GTACTGACAG ATAAAGCCTA GATGAGAGAG ATAGAGAGTG AGTACATGCA CACTCATGTG  
1740

CAAACCCACT CAGAGATGCA TTTGGAACAG TGCTACTGAA AGGCAGTAGT CATTTTCAAG  
1800

ACTGAATTCC AAACATGGTT TATTGGTGAG TTAGGAACAT GTAAGGCCAA GTACACTGAG  
1860

AGCCTTTTTG AAAGTAATTG AGTGGAAGT TGATGCCATT CTAAATCAAG GCATATCCAG  
1920

GTGGCCCGGT TTGAACTCAC TCCACTGTAC CCAGTCTCAA AGGCCAGGTT GCTAAGAAAC  
1980

CAGGAGTAAA AGAGTCAAGT GACCATCATT TCACCTGCTG CTTGCCCCCA ATAGTAGTCT  
2040

CTGTGAGGCC TTAGTGACCT CACCTAGGAA GTGATTTTTG AGCCCTTGTT TCAGGGCTGT  
2100

GGCCTCCCTG CTCTATCCTG AATAAAGCAG ACAGGTGTGC AGATTTTGGC CATGAAAGCA  
2160

TGGCTAATAG GGCCACAGTC CCTTTAAAGA AACATGGTTT GACTCTGGTT TTCTTGGGGG  
2220

09581423 413400

AAAATACCAC AATCACCGAT GCAAACATTG GAAGATTATT GAGAGCCCTA GAAAGCTGCT  
2280

GTGATCCCAG TAGAAAATAT GTCCCAGAAA TGTCATGAGA TTGCTGTGTG TTGCCTGGGA  
2340

CACAGATCAA GGGCCTATCT TGGAGAGCTG GGGACCAGCA GTCTGCCTGG AGGCCAGGGA  
2400

GCAGTGGCTG AGTAGCTCTG CCTTTGCTCT GGTCTATACC TTAAGAATGC CAAAGAATGA  
2460

ATTTCAACGC CTGCCTTTGG CACTCTGACT TAAAGTGCAA AAAGCTTCTG TGGCGAGGCA  
2520

TGCTATCATG GAATGAGACT GGCTTGCCCT AGGCTTAATG GATGGGCAGT CATTTTGCAG  
2580

AGGCTATGGG AAGAGGGTGA TAATAGAAGA GTGGCAGCTA TAGGAAATTA TCAACATACC  
2640

TTGGCCAGCA AGTTAGAGAA TCTGGCAATG GATGAACTGA AAGTGATGAA CTGGCAGGGA  
2700

TAACAAAGAA CCTAACATTT ATTAAGCACG TATTTATTAA CTGCTCAGTG TTTCATATTC  
2760

ATGCAAGTAT TCTCATTTTA CAGAGAAAGA AATTATGGCC CAGGGGGCTA AAGTAAACAA  
2820

CTCAAGGGCA CATAGAAAGT AAATAAAAGG ACTGTGATTT GAATCCAGGC CACTCTTAGC  
2880

CCATGCTGTT TTCCCTTTGC CACACTGTGG TAGGTGTTTG AACAGAGGCC ACATTACTAG  
2940

AGTTGGCATG ACTCTTGACT CTGCGCTGCC TAACAAAATA TTGAAAGGCA AACATTTGAA  
3000

GGAGGGAGGG GGTGCAGGTT CAGTTTATAT GGAAATGCAA ACTGGGCTGG AAGATATTCC  
3060

TGAGTTAGGC TTTCTCTTCA TATTCAGCTT GCACATTTGG TAATGTTTTT AAAATGATCA  
3120

TCTAATTTTA TTTTGTGAAG TGAAGGATTT GTGTTTTAGT TGGCAGTTGT TAAGTCCTTG  
3180

GCTTGCCATT TTTCAAAAG TAAAAAGGTC CTCACAGGTG TCTCCATACT TCGCCAAGGT  
3240

TGTAGCATGG GCAGTTTCAG TTTCAGCCTA AGAGACTGGT GACATCCACA AATGCAGTTT  
3300

TAGAAGCAGA AAAGGTCTTG GTGCCTCTGC AGTACTTGAT GTATTGGGGT CAAATCTCTA  
3360

CAAATTTTTT TGTGGTGATA GCAAATCAA GAGATGGCTT ACAAAAAGAA ATATTGAATT  
3420

09581423 112100

TTTATTTTGT AAGTTTTTGT TTTTAAAAG GTTGGGGGTG TTCAGCCACT GAGGGACAAA  
3480

ACTTAGCATC TAATTTCAAT TATAGTGTC TGCAGAGTAT TTCTAAAGTA ATTGGTTATC  
3540

ATGGGAAAGT ATTCTCTTTT CAAGAAGTTC TTTGATTCTG TAATAACTAG AACAAATAAA  
3600

GTAGTAAAAG AAGAAATAGT TCTGTGACTA GGAAAAAATT GCTTTTGAGA GAACATAGAT  
3660

CAATTATACT ACTTCTAAGG TAGCTGCAGA TAAGTGGCCT TGACACATTA CAAGCCTGGA  
3720

AAAAAACATC AGAAATAATA AAAAATTTCA GAGAGAATCA AGATACCTTT TTTTTTCTTT  
3780

TTTTTTTCTT TTTTTTATTA TACTCTAAGT TTTAGGGTAC ATGTGCACAT TGTGCAGGTT  
3840

AGTTACATAT GTATACATGT GCCATGCTGG TGCCTGTCAC CCACTAATGT GTCATCTAGC  
3900

ATTAGGTATA TCTCCCAGTG CTATCCCTCC CCCCTCCCCC GACCCACCA CAGTCCCCAG  
3960

AGTGTGATAT TCCCCTTCCT GTGTCCATGT GATCTCATTT TTCAATTCCC ACCTATGAGT  
4020

GAGAATATGC GGTGTTTGGT TTTTTGTTCT TGCATAGTT TACTGAGAAT GATGGTTTCC  
4080

AATTTTCATCC ATGTCCCTAC AAAGGATATG AACTCATCAT TTTTATGGC TGCATAGTAT  
4140

TCCATGGTGT ATATGTGCCA CATTTTCTTA ATCCAGTCTA TCATTGTTGG ACATTGGGGT  
4200

TGGTTCCAAG TCTTTGCTAT TGTGAATAGT GCCGCAATAA ACATACGTGT GCATGTGTCT  
4260

TTATAGCAGC ATGATTTATA CTCATTTGGG TATATACCCA GTAATGGGAT GGCTGGGTCA  
4320

AATGGTATTT CTAGTTCTAG ATCCCTGAGG AATCGCCACA CTGACTTCCA CAATGGTTGA  
4380

ACTAGTTTAC AGTCCAACCA ACAGTGTAAG AGTGTTTCTA TTTCTCCGCA TCCTCTCCAG  
4440

CACCTGTTGT TTCCTGACTT TTTAATGATT GCCATTCTAA CTGGTGTTGA GATGATATCT  
4500

CATA  
4504

- (i) SEQUENCE CHARACTERISTICS:  
    (A) LENGTH: 3101 base pairs  
    (B) TYPE: nucleic acid  
    (C) STRANDEDNESS: single  
    (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: cDNA
- (vi) ORIGINAL SOURCE:  
    (A) ORGANISM: Homo sapiens
- (ix) FEATURE:  
    (A) NAME/KEY: CDS  
    (B) LOCATION: 639..3047
- (ix) FEATURE:  
    (A) NAME/KEY: oligonucleotide  
    (B) LOCATION: 152..172
- (ix) FEATURE:  
    (A) NAME/KEY: oligonucleotide  
    (B) LOCATION: 586..606
- (ix) FEATURE:  
    (A) NAME/KEY: oligonucleotide  
    (B) LOCATION: 641..663

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 26:

TGTGCGTCGC GCTCTCGCCC TCCTCTTCCC GCTGCAGTGT CTATGGAGCG AGGCTACGTT  
60

TCATTGCCGC CCTGGCTTAA CCCTTCCGGC GCCTAAAAGG ACGGCCGGCC GGCCGGTCCC  
120

TTGCACCAGG AAGAAGTCTT AGCAGCCAGC GGGCCCTGGT CAGGAAACTC TAAGGTACAA  
180

GGAAAACAGT TGAGGAAGGA GCCAGAGCGC TCCGGTTTGG TCCTCGGGCT TCGCTGGGGC  
240

GGGGCGCAGG CGTTGGCTTT AAGAAAGGGG AGGGGACAGT GCAATCCGGG TTGCCCGCGG  
300

ATTCGGCCAA GGAATCTTCC GCTCGCTCCG GAGCGAGGAG CCTGTAAAGA GGCTGTTCCC  
360

AGCTCCAGCT CTAACCTCGC CTACACCTTG GCGGGGCCCA ATGTCACGTT TGCAATTGCT  
420

CAGGAAGGAT CCGGCCCCGTC TCCGGAGGCA AGTCGGGCTG CGGTTTTTGC TGCTTATCTG  
480

GGAAGGCGAT GCCTAAGGGA CATGCTGCTT GCTAGGCAGC ACCCTGCCGG GATCCGACTG  
540

CGATAGTTAG CTCTCCCTGG CCCTGAAGCC ATCGCCGGGG CGCCTGTTCT CTGTCCGGAC  
600

00727-2279550

CAGCCAGCGC TCCTCAGGAG TCTCACTGAA ACAGAACCAT GGGTCATCCC CCGCTGGAGT  
660

TCAGCGACTG CTACCTGGAC AGCCCCGATT TCCGCGAGAG GCTCAAGTGT TATGAGCAGG  
720

AACTGGAGAG GACCAACAAA TTCATCAAAG ACGTAATCAA AGACGGCAAC GCGCTTATCA  
780

GCGCTATGAG AAATTATTCT TCTGCTGTTT AGAAATTTTC CCAGACGCTG CAGTCATTTT  
840

AGTTTGATTT CATTGGAGAC ACTCTGACTG ATGATGAAAT TAACATCGCT GAATCCTTCA  
900

AGGAATTTGC TGAATGCTC AACGAGGTAG AAAATGAGAG GATGATGATG GTACACAATG  
960

CTAGTGATTT GCTGATTAAA CCCTTGGAAT ATTTCCGGAA GGAACAAATA GGCTTCACCA  
1020

AGGAGCGGAA AAAGAAATTT GAAAAGGATG GTGAGAGGTT TTATTCTTTA CTGGATCGGC  
1080

ACTTACACCT GTCTTCAAAA AAGAAAGAAT CTCAGTTACA AGAGGCAGAC CTACAGGTGG  
1140

ACAAGGAGAG GCACAATTTT TTCGAGTCCT CTCTTGATTA TGTTTATCAA ATCCAGGAAG  
1200

TTCAGGAGTC CAAGAAGTTC AATATTGTGG AGCCTGTCTT GGCCTTTCTT CATAGTCTGT  
1260

TCATTTCTAA CAGCCTGACT GTGGAGCTCA CACAGGATTT CCTCCCATAC AAACAACAGC  
1320

TCCAACCTCAG TTTACAGAAT ACAAGAAATC ATTTCTCCAG TACCCGGGAA GAGATGGAAG  
1380

AACTTAAGAA AAGGATGAAA GAAGCTCCCC AGACATGCAA ACTTCCAGGA CAGCCAACTA  
1440

TTGAAGGCTA TCTCTATACA CAAGAGAAAT GGGCTTTAGG AATATCCTGG GTGAAATACT  
1500

ATTGCCAGTA TGAGAAAGAG ACCAAAACAC TGACCATGAC GCCTATGGAG CAGAAGCCAG  
1560

GTGCTAAGCA GGGGCCCTTG GACTTAACAC TGAAGTACTG TGTGAGAAGG AAGACGGAGT  
1620

CTATCGACAA GAGGTTCTGT TTTGACATAG AACTAATGA AAGGCCAGGA ACCATCACTC  
1680

TGCAGGCCCT TTCAGAAGCT AACAGAAGGC TATGGATGGA AGCCATGGAT GGGAAAGAAC  
1740

CTATCTACCA CAGCCCTATA ACAAACAGC AAGAAATGGA GCTAAATGAA GTGGGCTTCA  
1800

0956443.1.21.00



AGTTTGTCTAG GAAGTGCATC AATATTATTG AGACCAAAGG GATCAAGACA GAAGGGTTGT  
1860

ACCGCACTGT GGGCAGCAAT ATTCAGGTTT AGAAGCTGCT GAATGCCTTT TTTGATCCTA  
1920

AATGCCCAGG AGATGTTGAT TTTCATAATA GTGACTGGGA CATTAGACA ATCACCAGCT  
1980

CCTTGAAATT CTACCTCAGG AATCTTTCTG AACCTGTCAT GACCTATAGA CTTCACAAAG  
2040

AGCTGGTCTC TGCTGCCAAG TCTGACAACC TGGATTACCG CCTAGGAGCT ATTCACTCCC  
2100

TGGTATATAA GCTACCAGAA AAGAACCGAG AGATGCTGGA ACTTCTGATA AGACACTTGG  
2160

TCAATGTGTG TGAGCACAGC AAAGAGAATC TTATGACCCC CTCCAACATG GGAGTAATCT  
2220

TTGGGCCCCAC CCTGATGAGA GCTCAGGAGG AACTGTGGC CGCCATGATG AACATCAAAT  
2280

TCCAGAACAT AGTGGTGGAA ATACTAATCG AGCACTTTGG CAAGATCTAT TTAGGTCCAC  
2340

CTGAGGAAAG CGCTGCACCG CCAGTGCCTC CGCCTCGGGT GACAGCAAGA AGGCACAAAC  
2400

CAATCACGAT TTCAAAGCGC TTGCTGCGAG AAAGGACGGT TTTCTATACT TCTTCCCTGG  
2460

ATGAAAGCGA AGATGAAATC CAACATCAAA CACCGAATGG TACTATCACC AGCAGCATAG  
2520

AACCCCCCAA GCCACCACAA CACCCCAAAC TACCTATTCA GAGGAGTGGG GAAACTGATC  
2580

CTGGGAGGAA GTCCCCAAGC AGGCCTATTT TGGATGGCAA GTTGGAGCCC TGCCAGAGG  
2640

TGGACGTGGG GAAGTTGGTG TCTAGGCTGC AGGATGGAGG GACCAAGATC ACCCCAAAGG  
2700

CCACCAATGG ACCCATGCCA GGCTCTGGGC CCACCAAGAC CCCCTCTTTC CACATAAAGA  
2760

GACCAGCTCC CCGGCCCTG GCCCACCACA AGGAGGGGGA TGCTGACAGT TTCAGCAAAG  
2820

TGCGGCCTCC AGGAGAAAAG CCAACCATCA TCCGCCCCC AGTGAGGCCC CCAGATCCTC  
2880

CCTGCCGGGC AGCTACTCCC CAAAAGCCAG AACCAAAGCC AGATATTGTG GCTGGCAATG  
2940

CGGGGGAAAT CACATCATCT GTGGTGGCTT CCAGGACCAG GTTTTTTGAA ACAGCTTCCC  
3000

09581422 443100

GGAAAACAGG AAGTTCTCAA GGCAGACTTC CTGGAGATGA AAGTTGAGGC TACAGGTTTT  
3060

AAAAGCCTTG GCCTCAGAGG ACCCTTTCCA GGTTCGAAA G  
3101

(2) INFORMATION FOR SEQ ID NO: 27:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 802 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(vi) ORIGINAL SOURCE:

- (A) ORGANISM: Homo sapiens

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 27:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Gly | His | Pro | Pro | Leu | Glu | Phe | Ser | Asp | Cys | Tyr | Leu | Asp | Ser | Pro | 1   | 5   | 10  | 15  |
| Asp | Phe | Arg | Glu | Arg | Leu | Lys | Cys | Tyr | Glu | Gln | Glu | Leu | Glu | Arg | Thr | 20  | 25  | 30  |     |
| Asn | Lys | Phe | Ile | Lys | Asp | Val | Ile | Lys | Asp | Gly | Asn | Ala | Leu | Ile | Ser | 35  | 40  | 45  |     |
| Ala | Met | Arg | Asn | Tyr | Ser | Ser | Ala | Val | Gln | Lys | Phe | Ser | Gln | Thr | Leu | 50  | 55  | 60  |     |
| Gln | Ser | Phe | Gln | Phe | Asp | Phe | Ile | Gly | Asp | Thr | Leu | Thr | Asp | Asp | Glu | 65  | 70  | 75  | 80  |
| Ile | Asn | Ile | Ala | Glu | Ser | Phe | Lys | Glu | Phe | Ala | Glu | Leu | Leu | Asn | Glu | 85  | 90  | 95  |     |
| Val | Glu | Asn | Glu | Arg | Met | Met | Met | Val | His | Asn | Ala | Ser | Asp | Leu | Leu | 100 | 105 | 110 |     |
| Ile | Lys | Pro | Leu | Glu | Asn | Phe | Arg | Lys | Glu | Gln | Ile | Gly | Phe | Thr | Lys | 115 | 120 | 125 |     |
| Glu | Arg | Lys | Lys | Lys | Phe | Glu | Lys | Asp | Gly | Glu | Arg | Phe | Tyr | Ser | Leu | 130 | 135 | 140 |     |
| Leu | Asp | Arg | His | Leu | His | Leu | Ser | Ser | Lys | Lys | Lys | Glu | Ser | Gln | Leu | 145 | 150 | 155 | 160 |
| Gln | Glu | Ala | Asp | Leu | Gln | Val | Asp | Lys | Glu | Arg | His | Asn | Phe | Phe | Glu | 165 | 170 | 175 |     |
| Ser | Ser | Leu | Asp | Tyr | Val | Tyr | Gln | Ile | Gln | Glu | Val | Gln | Glu | Ser | Lys | 180 | 185 | 190 |     |

Downloaded from www.wo-99-31230

Lys Phe Asn Ile Val Glu Pro Val Leu Ala Phe Leu His Ser Leu Phe  
 195 200 205  
 Ile Ser Asn Ser Leu Thr Val Glu Leu Thr Gln Asp Phe Leu Pro Tyr  
 210 215 220  
 Lys Gln Gln Leu Gln Leu Ser Leu Gln Asn Thr Arg Asn His Phe Ser  
 225 230 235 240  
 Ser Thr Arg Glu Glu Met Glu Glu Leu Lys Lys Arg Met Lys Glu Ala  
 245 250 255  
 Pro Gln Thr Cys Lys Leu Pro Gly Gln Pro Thr Ile Glu Gly Tyr Leu  
 260 265 270  
 Tyr Thr Gln Glu Lys Trp Ala Leu Gly Ile Ser Trp Val Lys Tyr Tyr  
 275 280 285  
 Cys Gln Tyr Glu Lys Glu Thr Lys Thr Leu Thr Met Thr Pro Met Glu  
 290 295 300  
 Gln Lys Pro Gly Ala Lys Gln Gly Pro Leu Asp Leu Thr Leu Lys Tyr  
 305 310 315 320  
 Cys Val Arg Arg Lys Thr Glu Ser Ile Asp Lys Arg Phe Cys Phe Asp  
 325 330 335  
 Ile Glu Thr Asn Glu Arg Pro Gly Thr Ile Thr Leu Gln Ala Leu Ser  
 340 345 350  
 Glu Ala Asn Arg Arg Leu Trp Met Glu Ala Met Asp Gly Lys Glu Pro  
 355 360 365  
 Ile Tyr His Ser Pro Ile Thr Lys Gln Gln Glu Met Glu Leu Asn Glu  
 370 375 380  
 Val Gly Phe Lys Phe Val Arg Lys Cys Ile Asn Ile Ile Glu Thr Lys  
 385 390 395 400  
 Gly Ile Lys Thr Glu Gly Leu Tyr Arg Thr Val Gly Ser Asn Ile Gln  
 405 410 415  
 Val Gln Lys Leu Leu Asn Ala Phe Phe Asp Pro Lys Cys Pro Gly Asp  
 420 425 430  
 Val Asp Phe His Asn Ser Asp Trp Asp Ile Lys Thr Ile Thr Ser Ser  
 435 440 445  
 Leu Lys Phe Tyr Leu Arg Asn Leu Ser Glu Pro Val Met Thr Tyr Arg  
 450 455 460  
 Leu His Lys Glu Leu Val Ser Ala Ala Lys Ser Asp Asn Leu Asp Tyr  
 465 470 475 480  
 Arg Leu Gly Ala Ile His Ser Leu Val Tyr Lys Leu Pro Glu Lys Asn  
 485 490 495  
 Arg Glu Met Leu Glu Leu Leu Ile Arg His Leu Val Asn Val Cys Glu  
 500 505 510

09584432 44100

His Ser Lys Glu Asn Leu Met Thr Pro Ser Asn Met Gly Val Ile Phe  
 515 520 525  
 Gly Pro Thr Leu Met Arg Ala Gln Glu Asp Thr Val Ala Ala Met Met  
 530 535 540  
 Asn Ile Lys Phe Gln Asn Ile Val Val Glu Ile Leu Ile Glu His Phe  
 545 550 555 560  
 Gly Lys Ile Tyr Leu Gly Pro Pro Glu Glu Ser Ala Ala Pro Pro Val  
 565 570 575  
 Pro Pro Pro Arg Val Thr Ala Arg Arg His Lys Pro Ile Thr Ile Ser  
 580 585 590  
 Lys Arg Leu Leu Arg Glu Arg Thr Val Phe Tyr Thr Ser Ser Leu Asp  
 595 600 605  
 Glu Ser Glu Asp Glu Ile Gln His Gln Thr Pro Asn Gly Thr Ile Thr  
 610 615 620  
 Ser Ser Ile Glu Pro Pro Lys Pro Pro Gln His Pro Lys Leu Pro Ile  
 625 630 635 640  
 Gln Arg Ser Gly Glu Thr Asp Pro Gly Arg Lys Ser Pro Ser Arg Pro  
 645 650 655  
 Ile Leu Asp Gly Lys Leu Glu Pro Cys Pro Glu Val Asp Val Gly Lys  
 660 665 670  
 Leu Val Ser Arg Leu Gln Asp Gly Gly Thr Lys Ile Thr Pro Lys Ala  
 675 680 685  
 Thr Asn Gly Pro Met Pro Gly Ser Gly Pro Thr Lys Thr Pro Ser Phe  
 690 695 700  
 His Ile Lys Arg Pro Ala Pro Arg Pro Leu Ala His His Lys Glu Gly  
 705 710 715 720  
 Asp Ala Asp Ser Phe Ser Lys Val Arg Pro Pro Gly Glu Lys Pro Thr  
 725 730 735  
 Ile Ile Arg Pro Pro Val Arg Pro Pro Asp Pro Pro Cys Arg Ala Ala  
 740 745 750  
 Thr Pro Gln Lys Pro Glu Pro Lys Pro Asp Ile Val Ala Gly Asn Ala  
 755 760 765  
 Gly Glu Ile Thr Ser Ser Val Val Ala Ser Arg Thr Arg Phe Phe Glu  
 770 775 780  
 Thr Ala Ser Arg Lys Thr Gly Ser Ser Gln Gly Arg Leu Pro Gly Asp  
 785 790 795 800  
 Glu Ser

00984422.143100